Behaviormetrika (2018) 45:317–334
https://doi.org/10.1007/s41237-018-0058-8

REVIEW PAPER

Propensity score methods for causal inference: an overview

Wei Pan[1] · Haiyan Bai[2]

Received: 29 January 2018 / Accepted: 18 July 2018 / Published online: 24 July 2018

Abstract

Propensity score methods are popular and effective statistical techniques for reducing selection bias in observational data to increase the validity of causal inference based on observational studies in behavioral and social science research. Some methodologists and statisticians have raised concerns about the rationale and applicability of propensity score methods. In this review, we addressed these concerns by reviewing the development history and the assumptions of propensity score methods, followed by the fundamental techniques of and available software packages for propensity score methods. We especially discussed the issues in and debates about the use of propensity score methods. This review provides beneficial information about propensity score methods from the historical point of view and helps researchers to select appropriate propensity score methods for their observational studies.

Keywords Propensity scores · Propensity score methods · Propensity score analysis · Propensity score matching · Subclassification · IPTW

1 Introduction

Randomized controlled trials (RCTs) are regarded as the best research design for causal inference. However, RCTs are not always feasible in behavioral and social science research due to practical or ethical barriers. Consequently, non-RCTs (or observational studies) are often used as an alternative. Unfortunately, the validity of such studies is often called into question because of potential selection bias in observational data (Rosenbaum 2010; Shadish et al. 2002). To increase the validity of observational studies, multiple strategies have been developed to deal with selection bias.

Communicated by Takahiro Hoshino.

Wei Pan
wei.pan@duke.edu

[1] Duke University, DUMC 3322, 307 Trent Dr, Durham, NC 27710, USA

[2] University of Central Florida, PO Box 161250, Orlando, FL 32816, USA

One of those strategies, propensity score methods, is most popular. This popularity stems from the methods' properties that mimic those of RCTs (Bai 2011b; Pan and Bai 2016b; Rubin 2008). While many researchers tout the advantages of propensity score methods, some methodologists and statisticians raise concerns about the rationale and applicability of propensity score methods (Pearl 2010; King and Nielsen 2016). In this review, we address these concerns by reviewing the development history and the assumptions of propensity score methods, followed by the fundamental techniques of and software packages for propensity score methods. We also discussed the issues in and debates about the use of propensity score methods.

2 Development history of propensity score methods

In their seminal work of propensity score methods, Rosenbaum and Rubin (1983b) introduced the fundamental concept of a propensity score and its basic applications in observational studies. They defined a propensity score as the conditional probability of assignment to a particular treatment given observed covariates. They adopted the causal framework for treatment effect estimation from previous eminent literature (Fisher 1951; Hamilton 1979; Kempthorne 1952; Rubin 1974, 1978). Based on both the large- and small-sample theories, Rosenbaum and Rubin (1983b) proved that the adjustment using propensity scores calculated from all observed covariates is sufficient to remove selection bias in observational data.

The key concept behind propensity score methods is that they can be used to balance the distributions of covariates between the treatment and control groups. This logic is based on Rubin's (1978) causal effect theory and employs the " ignorability assumption". This simply means that the choice to assign a subject to the control or treatment group is effectively random when conditioned on observable characteristics, and missing data can be treated as occurring randomly as well. It is a fundamental assumption of propensity score methods.

In Rosenbaum and Rubin's (1983b) study, they expounded on Cochran and Rubin's (1973) previous work on the use of a single normally distributed covariate. They tested the use of propensity scores calculated from multiple covariates to adjust the unbalanced distributions of covariates between the treatment and control groups, and made three proposals for standard applications of propensity scores: (a) propensity score matching, (b) subclassification based on propensity scores, and (c) multivariate covariate adjustment.

Propensity score matching creates a matched set of subjects in the control group to those in the treatment group with similar propensity scores. The goal is to mimic random selection and eliminate bias in observational data. They articulated four reasons for the advantages of propensity score matching over model-based alternative adjustment on random samples: (a) propensity score matching allows researchers to easily analyze matched pairs and adjust for confounding variables; (b) the variance of the estimate of the average treatment effect is smaller in the matched sample than in random samples because the distributions of the covariates in the matched sample are similar; (c) model-based adjustment on matched samples is more robust than the adjustment on random samples; and (d) small sample sizes do not allow control of multiple covariates with model-based methods, but propensity score matching does.

In a later study, Rosenbaum and Rubin (1985) developed three different propensity score matching techniques, with an emphasis on multivariate matching: (a) nearest available matching on the estimated propensity scores, (b) Mahalanobis metric matching including propensity scores, and (c) nearest available Mahalanobis metric matching within a caliper defined by propensity scores.

Nearest available matching (nearest neighbor matching or greedy matching) was initially defined as matching the treated and control subjects on their closest propensity scores without replacement, with the treated and control subjects randomly ordered. Later, it was found that propensity score matching would create different matched pairs if the subjects were not chosen randomly, but in order, such as from the smallest propensity score to the largest or from the largest to the smallest. Furthermore, matching with and without replacement also created significantly different matched pairs.

Mahalanobis metric matching with propensity scores is a procedure that includes the estimated propensity scores in the calculation of Mahalanobis distance and then to match the sample without replacement on the Mahalanobis distance with the treated and control subjects randomly ordered. A variation of Mahalanobis metric matching with propensity scores is the nearest available Mahalanobis metric matching within a caliper that creates matched pairs of treated and control subjects using Mahalanobis distance within a caliper band defined by propensity scores so as to control the difference between matched pairs.

While Rosenbaum and Rubin (1985) clearly described the procedures of propensity score matching in their study and favored caliper matching, they did not examine significant issues with propensity score matching, such as ranking the order of subjects on propensity scores, the size of calipers, and matching with or without replacement. Further, their discussion of propensity score matching was limited to only one sample.

Meanwhile, Rosenbaum and Rubin (1984) developed another propensity score method to reduce selection bias in observational studies using subclassification on propensity scores. Their simulation study showed that subclasses formed using propensity scores can balance all the covariates, and that five subclasses could remove up to 90% of the bias for each of the covariates.

In 1989, Rosenbaum introduced the *optimal matching* algorithm based on network flow theory, a departure from the previous *greedy matching* procedures, such as nearest neighbor matching and caliper matching. Optimal matching generally identifies matched pairs that minimize the total distance in propensity scores between the treatment and control groups. In optimal matching, one treated subject can be matched with multiple control subjects; therefore, it has the advantage of keeping all the control subjects. Rosenbaum (1989) claimed that optimal matching is often better than greedy matching, but this should not be taken as an absolute conclusion because the performance of different matching techniques can be data specific.

Following Rosenbaum and Rubin's pioneer works in 1970s and 1980s, other techniques of propensity score methods were also developed. Recent developments have focused mainly on the improvement of propensity score estimation accuracy and matching quality. Improving propensity score estimates involves: (a) the classification and regression tree procedure to examine each predictor variable for creating two distinctly different samples (Lemon et al. 2003), (b) boosted regression using regression tree to derive propensity scores (McCaffrey et al. 2004), and (c) bootstrapping propensity scores to account for sampling errors (Bai 2013).

Improvements to matching quality include: (a) *kernel matching* that utilizes weighted regression in matching procedures (Heckman et al. 1997), (b) *full matching* that incorporates optimal matching with replacement of the subjects from both the treatment and control groups (Hansen 2004), (c) *genetic matching* that is a nonparametric procedure using a search algorithm to determine the weight of each covariate for maximizing the balance of observed potential confounders across the matched treated and control subjects (Diamond and Sekhon 2013), and (d) *interval matching* that matches subjects based on confidence intervals other than point estimates of propensity scores to accommodate estimation errors in propensity scores (Pan and Bai 2015b).

In the meantime, Hirano and Imbens (2001) also proposed model-based direct adjustment using propensity score weighting, which is defined as the inverse probability of treatment weighting (IPTW) using propensity scores. IPTW has grown in popularity because of its capacity to deal with complex data. Austin and Stuart (2015) described IPTW in detail and pointed out that to use the adjustment correctly, researchers must examine whether weighting balances covariates between the treatment and control groups.

3 Assumptions of propensity score methods

Like all other statistical methods, assumptions are required for applying propensity score methods. These assumptions are the ignorable treatment assignment assumption, the stable unit treatment value assumption, and sufficient common support.

3.1 The ignorable treatment assignment assumption

Suppose each subject (or unit) i ($i=1, \ldots, N$) has a treatment condition Z_i (1 = treatment or 0 = control), an outcome Y_{1i} (potential outcome for treated unit) or Y_{0i} (potential outcome for control unit), and a covariate value vector $\mathbf{X}_i=(X_{i1},\ldots, X_{iK})^{\mathrm{T}}$, where N is the number of subjects and K is the number of covariates. The ignorable treatment assignment assumption requires that assignment to the treatment or control group is independent of both outcomes after accounting for observed covariates: $(Y_{1i}, Y_{0i}) \perp Z_i \mid \mathbf{X}_i$. Under this assumption, if the distributions of propensity scores are balanced between the treatment and control groups, the distributions of the covariates used for obtaining propensity scores are also balanced. That is, $(Y_{1i}, Y_{0i}) \perp Z_i \mid \mathbf{X}_i \Rightarrow (Y_{1i}, Y_{0i}) \perp Z_i \mid p(\mathbf{X}_i)$, where $p(\mathbf{X}_i)=Pr(Z_i=1 \mid \mathbf{X}_i)$ defined as a propensity score. Therefore, one can assume that selection bias can be removed or significantly reduced after propensity score adjustments if no confounding variables are left unmeasured. This is the foundation of propensity score theory. Therefore, all influential covariates associated with the treatment estimation and outcomes should be included in the propensity score estimation model. In reality, hidden bias often exists because we are unable to include unobserved covariates in the propensity score estimation model; therefore, treatment effect estimation will be affected.

One way to address this unobserved confounding issue is to conduct sensitivity analysis. Sensitivity analysis assesses the impact of an unobserved confounding variable under certain assumptions made by the researcher. This technique helps researchers better understand the limitation of propensity score methods due to unobserved confounding, and more and more researchers have been conducting sensitivity analysis (Groenwold et al. 2010; Schneeweiss 2006; Lin et al. 1998; Robins et al. 2000b; Rosenbaum and Rubin 1983a). There are several variants of sensitivity analysis with specific techniques, such as marginal structural models (Robins et al. 2000a), linear programming (MacLehose et al. 2005), Bayesian sensitivity analysis (Greenland 2005), external adjustment (Huesch 2013), propensity score-based approach (Li et al. 2011), and the robustness index (Pan and Bai 2016a). Among these techniques, Rosenbaum and Rubin's (1983a) is the most frequently used with a ready-to-implement statistical package in R (Keele 2015).

3.2 The stable unit treatment value assumption

The stable unit treatment value assumption (SUTVA) requires that the treatment effect for each subject be independent of other subjects' responses; thus, the treatment for each subject is stable or the same (Rosenbaum and Rubin 1983b). To apply propensity score methods, such as propensity score matching, SUTVA assumes that: (a) the potential outcome of the selected subjects in the treatment group should not be affected by the treatment status of other subjects in the study groups and (b) each subject receives the same amount of treatment as the others in the treatment group who were selected through propensity score matching.

In practice, SUTVA can be easily violated if subjects in the treatment group interact with subjects in the control group. For example, a subject in the treatment group of a weight loss exercise program may like to share his or her experience of the treatment with his or her friends who happen to be in the control group. Such information sharing between the treatment and control groups can make the potential outcomes dependent on each other, or the treatment is not given consistently to each subject in the treatment condition, which makes the treatment effect unstable. Therefore, an appropriate research design followed by a rigorous procedure for complying with the research protocol is needed to reduce the likelihood of violating SUTVA.

3.3 Sufficient common support

The third assumption is about common support (or overlap) between the distributions of propensity scores for the treatment and control groups. It implies that the propensity scores of the two groups should overlap. This allows researchers to make a reasonable (or unbiased) comparison between the two groups. Common support can be improved by having more variability in propensity scores in the control group than the treatment group (Pan and Bai 2015a). In practice, common support improvement can be achieved by having proportionally more participants in the control group who can be matched to those in the treatment group.

There are several ways of checking common support. First, we can make a visual inspection of propensity score distributions with histograms or density graphs of propensity score distributions. Second, we can use hypothesis testing, such as Kolmogorov–Smirnov test, to determine if propensity score distributions are significantly different from each other. Third, we can compute the standardized difference score: $d = \frac{M_t - M_c}{s_p}$ to compare the means of the propensity scores for the treatment (M_t) and control (M_c) groups, where s_p is the pooled standard deviation of the propensity scores. A small standardized difference score (e.g., $d < 0.5$) indicates sufficient common support. Last, we can trim the minimum and maximum values of propensity scores in each group by removing the subjects whose propensity scores are smaller than the minimum or larger than the maximum in the opposite group (Caliendo and Kopeinig 2008; Pan and Bai 2015a; Smith and Todd 2005).

4 Fundamental techniques of propensity score methods

The fundamental techniques of propensity score methods can be generally classified into five categories: propensity score matching (e.g., nearest neighbor matching, clipper matching, Mahalanobis matching with propensity scores) (Rosenbaum and Rubin 1985), subclassification on propensity scores (Rosenbaum and Rubin 1984), propensity score weighting (Hirano and Imbens 2001), covariate adjustment with propensity scores, and doubly robust estimation. To better understand the fundamental techniques, we start discussing these topics with the underlying theoretical framework of causal inference.

4.1 Causal inference and propensity score methods

In the counterfactual framework for causal inference, the quantity of interest is the treatment effect for each subject i, which is defined as $\Delta_i = Y_{1i} - Y_{0i}$ (Holland 1986; Rubin 1974; Winship and Morgan 1999), where Y_{1i} is the potential outcome for an individual subject exposed to the treatment, while Y_{0i} is the potential outcome for the same individual without receiving any treatment at the same time. Unfortunately, for each subject i, only Y_{1i} or Y_{0i}, is observable, but not both, at the same time because the same subject cannot be simultaneously assigned to both the treatment or control conditions. Alternatively, one can estimate the *average treatment effect* (ATE) for the population, which is defined as $\text{ATE} = E(Y_1 - Y_0) = E(Y_1) - E(Y_0)$, where $E(Y_1)$ is the expected value of outcome Y_1 for all the subjects in the treatment group and $E(Y_0)$ is the expected value of Y_0 for all the subjects in the control group. In RCTs, ATE is an unbiased estimate of the treatment effect of Δ_i because the treatment group does not, on average, differ systematically from the control group on their observed and unobserved background characteristics, due to randomization. In observational studies, treatment effect estimation could be biased because the treatment and control groups may not be comparable due to the potential group selection bias without randomization.

Selection bias can be overt, hidden, or both (Rosenbaum 2010). Fortunately, propensity score methods set forth by Rosenbaum and Rubin (1983b) can reduce overt bias in observational studies by balancing the distributions of observed characteristics (or covariates) between the treatment and control groups. Therefore, propensity score methods allow one to obtain an unbiased estimate of ATE from observational studies under the assumption of the ignorable treatment assignment assumption.

ATE is not always the quantity of interest (Heckman et al. 1997; Rubin 1977). For instance, one may be interested in the treatment effect of a smoking cessation program for smokers who volunteered to participate in the program, not necessarily for all people in the population. In this case, one wants to estimate the *average treatment effect for the treated* (ATT) for the population, which is defined as $\text{ATT} = E(Y_1 - Y_0| Z = 1) = E(Y_1| Z = 1) - E(Y_0| Z = 1)$. This still encounters the counterfactual problem that one can only observe the average treatment effect Y_1 for people who receives the treatment, but it is not observable for the effect Y_0 for the participants had they not been treated. To deal with this problem, one can analyze matched data on propensity scores. The matched subjects in the control group have similar probabilities of $Z = 1$ to these of the corresponding subjects in the treatment group and, therefore, propensity score methods allow one to estimate ATT.

4.2 Covariates selection and propensity score estimation

Before implementing the techniques of propensity score methods, we first need to estimate propensity scores from given covariates. It is essential to include all influential covariates in the propensity estimation model so that the propensity scores used for balancing the distributions of the covariates will be accurate. While we can use statistics, such as correlations, to guide covariates selection, covariates should be selected based on the theory or existing literature about the relationships of the covariates to the outcome variables and treatment assignment conditions (Rubin 2001).

In general, there are three types of relationships of covariates to the treatment conditions and the outcome variables. First, the best covariates to be selected are those related to both the treatment conditions and outcome variables, since the relationships indicate that the covariates may both alter the treatment and influence the treatment effect. Second, if a covariate is associated with the outcome variable, but not with the treatment, it may still change the outcome estimation; therefore, it needs to be included in the propensity score model (Rubin and Thomas 1996; Brookhart et al. 2006). Third, if a variable is related to the treatment, but not outcome, the decision to include or exclude that variable in the propensity score model depends on the direction of the relationship between the potential covariate and the treatment condition (Brookhart et al. 2006). If the covariate *has an impact on the treatment*, it *should* be used for the propensity score estimation, as it could alter the treatment.

However, if the covariate is associated with the treatment conditions, but *does not have an impact on the treatment* (nor is it related to the outcome variable), it *should not be* included in the model because this type of variable will not affect treatment effect.

After establishing a set of influential covariates, the next step is to estimate propensity scores. Rosenbaum and Rubin (1983b) first recommended using logistic regression or discriminant analysis to estimate propensity scores. It is worth noting that in logistic regression or discriminant analysis, we only need to model an assignment probability given covariates without assuming any functional form of the distributions of the probability. That is, propensity score estimation is semiparametric and thus robust. Nevertheless, more advanced techniques, such as classification and regression trees, neural networks, and bootstrap techniques, were later adopted for obtaining more accurate propensity score estimation (Westreich et al. 2010). Propensity scores can be obtained by running these models using statistical software, such as R, SAS, or STATA. There are many statistical packages for implementation of propensity score methods. It is common for these statistical packages to include propensity score estimation in the propensity score adjustment procedures. In the next section, we will introduce commonly used propensity score methods.

4.3 Propensity score matching

Before the inception of propensity score matching, *exact matching* was the traditional matching technique used in quasi-experimental designs for matching on a few categorical variables. Another traditional matching technique, *Mahalanobis matching*, was used for matching on multivariate continuous variables (Rubin 1980). Then, two sets of propensity score matching algorithms, greedy matching and complex matching, followed. Bai (2013: Fig. 1) defined a propensity score matching typology that depicts the developmental relationships among the propensity score matching techniques.

Nearest neighbor matching (Rosenbaum and Rubin 1985) is the foundation of all propensity score matching techniques. It is based on the greedy matching algorithm that matches each subject i in the treatment group with a subject j in the control group by the smallest absolute distance between their propensity scores: $d_i = \min_j |p(\mathbf{X}_i) - p(\mathbf{X}_j)|$. To conduct nearest neighbor matching without replacement, we need to decide how to rank the subjects, namely randomly, largest to smallest, or smallest to largest, based on their propensity scores. For matching with replacement, the ranking order is not needed because the same control subject can be used multiple times. An alternative to nearest neighbor matching is *caliper matching* (Cochran and Rubin 1973), which matches each subject i in the treatment group with a subject j in the control group within a prespecified caliper band b: $d_i = \min_j \{ |p(\mathbf{X}_i) - p(\mathbf{X}_j)| < b \}$, to reduce the risk of bad matches when the distance of the propensity scores between the matched pairs is too great. Based on Cochran and Rubin's (1973) study, Rosenbaum and Rubin (1985) recommended that the prespecified caliper band should be less than or equal to a quarter of the standard deviation of the propensity scores; such a caliper will remove up to 90% of selection bias. In practice, the caliper bandwidth can also be defined by the researcher to make better matched pairs; however, it is usually challenging for researchers to select a reasonable data-specific caliper or to detect the tolerance level on the maximum propensity score distance. A variant of caliper matching is *radius matching* (Dehejia and Wahba 2002), which is a one-to-many matching with each subject i in the treatment group matched with multiple subjects in the control group within a prespecified caliper band: $d_{ij}=\{|p(\mathbf{X}_i) - p(\mathbf{X}_j)|<b\}$. Recently, Pan and Bai (2015b) extended caliper matching to *interval matching* that matches subjects based on confidence intervals in propensity scores to accommodate estimation errors of propensity scores. In other words, if the confidence intervals (CI) of propensity scores overlap: $\mathrm{CI}(p(\mathbf{X}_i)) \cap \mathrm{CI}(p(\mathbf{X}_j)) \neq \varnothing$, the two subjects are taken as a matched pair.

Another type of greedy matching is *Mahalanobis metric matching with propensity scores* (Rosenbaum and Rubin 1985), which matches each subject i in the treatment group with a subject j in the control group according to the closest Mahalanobis distance calculated on proximities of the variables: $d_i=\min_j\{D_{ij}\}$, where $D_{ij}=(\mathbf{V}_i - \mathbf{V}_j)\mathbf{S}^{-1}(\mathbf{V}_i - \mathbf{V}_j)^{\mathrm{T}}$, where $\mathbf{V}$ is a combined data matrix $\{\mathbf{X}, p(\mathbf{X})\}$ and $\mathbf{S}$ is the sample variance–covariance matrix of $\mathbf{V}$ for the control group. In practice, this matching method does not perform as well as other propensity score matching techniques such as nearest neighbor matching and caliper matching (Bai 2011a). *Mahalanobis caliper matching* (Guo et al. 2006) and *genetic matching* (Diamond and Sekhon 2013) are two variants of Mahalanobis metric matching with propensity scores. Mahalanobis caliper matching uses $d_i=\min_j\{D_{ij}<b\}$, where $D_{ij}=(\mathbf{X}_i - \mathbf{X}_j)\mathbf{S}^{-1}(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}$; and genetic matching uses the weighted Mahalanobis: $D_{ij}=(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{S}^{-1/2})^{\mathrm{T}}\mathbf{W}\mathbf{S}^{-1/2}(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}$ or $D_{ij}=(\mathbf{V}_i - \mathbf{V}_j)(\mathbf{S}^{-1/2})^{\mathrm{T}}\mathbf{W}\mathbf{S}^{-1/2}(\mathbf{V}_i - \mathbf{V}_j)^{\mathrm{T}}$, where $\mathbf{W}$ is a weighting matrix and $\mathbf{S}^{1/2}$ is the Cholesky decomposition of $\mathbf{S}$.

It is worth noting that after propensity score matching, treatment effect estimation on the matched data will only give us ATT as RCTs do. To estimate ATE, one needs to use the entire original data with propensity score weighting, covariate adjustment with propensity scores, or propensity score matching-related methods, as described below.

There are propensity score matching-related or complex matching methods, some of which do not strictly match individual subjects. For example, *subclassification* (or stratification) (Rosenbaum and Rubin 1984) classifies all the subjects in a sample into several strata based on the corresponding number of percentiles of propensity scores and then matches stratum by stratum instead of individual subjects. Cochran and Rubin (1973) observed that five strata would remove up to 90% of selection bias. *Optimal matching* (Rosenbaum 1989) is another type of complex matching with an algorithm that is strikingly different from that of greedy matching. In greedy matching, after a match is made, the matched pairs will not be reconsidered. Each pair of matched subjects is considered the best matched pair currently available, whereas in optimal matching, previously matched pairs can be reconsidered to achieve the overall minimal or optimal distance. Optimal matching is particularly useful when there are not many appropriate control subjects to be matched with the treated subjects. An extension of optimal matching is *full matching* (Hansen 2004), which is also considered a special case of subclassification. Full matching produces subclasses in an optimal way. A fully matched sample consists of matched subsets, in which each matched set can contain one treated subject and one or more control subjects, or one control subject and one or more treated subjects. Full matching is optimal because it minimizes a weighted average of the estimated distance measure between each treated subject and each control subject within each subclass. The last type of complex matching is *kernel matching* (or local linear matching) (Heckman et al. 1997). It combines matching and outcome analysis into one procedure with one-to-all matching by using nonparametric matching estimators to obtain weighted averages of all subjects in the control group for constructing the counterfactual outcome. A variant of kernel matching is *difference-in-differences matching*, which calculates the differences between the outcome of the treated subjects and the weighted average differences in outcome for the control subjects (Heckman et al. 1997).

4.4 Propensity score weighting

Propensity score weighting, such as IPTW (Hirano and Imbens 2001), is another useful propensity score method which combines propensity scores directly into treatment effect estimations. The IPTW estimator weights the observations of the dependent variable by the inverse of the propensity scores for balancing the treatment and control groups. This propensity score weighting approach becomes increasingly popular because of its capacity to deal with multiple data formats, such as nested data, longitudinal data, and multi-treatment data; and therefore, it has been greatly discussed and applied in the literature (Harder et al. 2010; McCaffrey et al. 2004; Schafer and Kang 2008; Stone and Tang 2013).

IPTW can be used to estimate both ATE and ATT. In ATE, weights are applied to both the treatment and control groups, whereas ATT only weights the control group. To estimate ATE, one first weights the observations for the treatment group by the inverse of the propensity score: $Y_{wt_i} = \frac{1}{p(\mathbf{X}_i)} Y_{t_i}$, and then weight the observations for the control group by the inverse of 1 minus the propensity score: $Y_{wc_i} = \frac{1}{1-p(X_i)} Y_{c_i}$, where Y_{wt_i} and Y_{wc_i} are the weighted observations of the dependent variable for each subject in the treatment group and each subject in the control group, respectively; Y_{t_i} and Y_{c_i} are the original observations for those subjects in the treatment and control groups, respectively. Then, the weighted observations are summed and divided by the total sample size ($N = n_\mathrm{t} + n_\mathrm{c}$, where n_t and n_c are the sample sizes of the treatment and control groups, respectively). Lastly, ATE is the difference between the two averages: $\mathrm{ATE} = \frac{1}{N}\left(\sum_{i=1}^{n_t} Y_{wt_i} - \sum_{i=1}^{n_c} Y_{wc_i}\right)$. As opposed to ATE, ATT does not weight the observations in the treatment group but only the control group by the ratio of the propensity score to the inverse of 1 minus the propensity score: $Y_{wc_i} = \frac{p(\mathbf{X}_i)}{1-p(\mathbf{X}_i)} Y_{c_i}$; and thus, ATT is computed as follows: $\mathrm{ATT} = \frac{1}{n_t} \sum_{i=1}^{n_t} Y_{t_i} - \frac{1}{n_c} \sum_{i=1}^{n_c} Y_{wc_i}$.

4.5 Covariate adjustment with propensity scores

To control the confounding factors in observational studies, analysis of covariance is commonly used to partial out the effects of confounding effect on the treatment effect by including influential covariates in the statistical model. These covariates are assumed to have effects on outcomes and, therefore, confound the treatment effect estimation. Even though traditional covariate analyses *can* control for confounding factors to some extent, it only decomposes the variance in the outcome into variance explained by the covariates, variance explained by the treatment conditions, and residual variance. However, it is difficult to determine if the covariate analysis model is correctly specified the relationships between treatment selection and baseline covariates to the outcome (Austin 2011). Therefore, covariate analysis does not model the selection bias since it does not analyze the confounding effect directly resulting from the unbalanced covariates distributions for the treatment and control groups. Therefore, the simplest method for using propensity scores is to include the propensity scores in the regression model to adjust the contributions of the covariates to the treatment effect based on the composite score of all the covariates to account for the probability of the individual subject to be assigned to one of the treatment conditions, commonly to the treatment condition as defined in the propensity score method.

Covariate adjustment with propensity scores is usually applied to the entire original data to obtain ATE, which is the estimated regression coefficient β_1 from the following multiple regression model with propensity score adjustment on the entire original data: $Y_i = \beta_0 + \beta_1 Z_i + \beta_2 p(\mathbf{X}_i) + \beta_3 Z_i p(\mathbf{X}_i) + \varepsilon_i$, where Z_i is the treatment condition, $p(\mathbf{X}_i)$ is the propensity score, and $Z_i p(\mathbf{X}_i)$ is the interaction between the treatment condition and the propensity score.

4.6 Doubly robust methods

In practice, these propensity score methods may not sufficiently reduce selection bias when propensity score estimation model is misspecified (Schafer and Kang 2008). Model misspecification may occur if any influential covariates are not included in the propensity score estimation model or if there are any misspecified forms of covariates, such as interaction effect, higher-order terms, or non-linear trends. Model misspecification can happen not only for the propensity score estimation model, but also in the outcome regression model. In this situation, using doubly robust methods will increase the accuracy of outcome estimation after propensity score adjustments. Doubly robust estimation incorporates outcome regression model and propensity score model in treatment effect estimation, which is robust to one model misspecification (either regression model or propensity score model) (Bang and Robins 2005; Li et al. 2017). Doubly robust procedures were found to reduce more bias than just using one propensity score procedure alone (Shadish et al. 2008). Therefore, doubly robust estimation is increasingly used when implementing propensity score methods.

Doubly robust procedures can be used with many types of propensity score adjustment methods. Schafer and Kang (2008) suggest using a doubly robust procedure, in which the individual covariates are still included in the treatment effect estimation model after the propensity scores adjustments. Imai and Ratkovic (2014) proposed covariate balancing propensity score which exploits the dual characteristics of the propensity score as a covariate balancing score and the conditional probability of treatment assignment using generalized method-of-moments or empirical likelihood framework. The method is found to improve the performance of propensity score weighting, and can be extended to non-binary treatment conditions and longitudinal data, and generalizing experimental and instrumental variable estimates (Imai and Ratkovic 2014). While it is worth noting that doubly robust procedure is appealing, it can still result in biased estimate if both regression model and propensity score model are misspecified (Funk et al. 2001; Li et al. 2017).

4.7 Evaluation of covariate balance

It is important to evaluate covariate balance before and after propensity score matching. Prior to evaluating covariate balance, common support of propensity score distributions should be assessed. As discussed in Sect. 3.3, common support should be sufficient in propensity score matching to create good matched pairs, which ensures the matched subjects in the control group are similar to those in the treatment group in probability of assignment to the treatment group. Thus, propensity score matching with sufficient common support can approximate random assignment similar to that of RCTs. Although the existing literature (e.g., Caliendo and Kopeinig 2008; Heckman et al. 1997; Rubin 2001) discusses the importance of sufficient comment support, most studies do not include a standard for how much common support is sufficient for propensity score matching. Based on Bai (2015), we recommend that if 75% of the propensity scores overlap, this may provide a better sample pool for finding matching pairs.

Before implementing propensity score matching, it is essential to evaluate covariate balance to understand the status of selection bias. If the treatment and control groups are well balanced on all the covariates, it means that there is no selection bias and no need to conduct propensity score matching, subclassification, or weighting procedures. After propensity score matching, it is important to evaluate covariate balance again to see if selection bias is sufficiently reduced. If not, further model-based covariate adjustment should be considered.

There are three criteria for evaluating covariance balance to see if selection bias exists in any of the covariates. First, *selection bias* in the *k*th covariate is defined as the mean difference in the covariate between the treatment and control groups: $B_k = M_{t_k} - M_{c_k}$, where M_{t_k} and M_{c_k} are the means of the covariate for the treatment and control groups, respectively. Intuitively, an independent-sample *t* test for continuous covariates or chi-square test for categorical covariates could be readily applied to test the selection bias. However, researchers should be cautious about using significant tests as the only means by which to evaluate covariate balance. The aim of evaluating covariate balance is not to test for sample differences that may be affected by factors such as sample size and variance instead of covariate imbalance (Pan and Bai 2016b). Second, we can examine the *standardized bias* (*SB*) for a covariate: $SB_k = \frac{B_k}{s_{\mathrm{p}_k}} = \frac{M_{\mathrm{t}_k} - M_{\mathrm{c}_k}}{s_{\mathrm{p}_k}}$, where s_{p_k} is the pooled standard deviation of the covariate (Rosenbaum and Rubin 1985). If the absolute value of SB_k is less than 0.05 (or 5%), the matching method is considered effective in balancing the covariate (Caliendo and Kopeinig 2008). The third criterion is the *percent bias reduction* (PBR) on the covariate: $PBR_k = \frac{B_{k,\text{ before matching}} - B_{k,\text{ after matching}}}{B_{k,\text{ before matching}}}$, with a PBR_k larger than 0.80 (or 80%) indicating an effective bias reduction (Cochran and Rubin 1973).

In addition, graphics are a means of evaluating and visualizing covariate balance, such as *Q–Q* plot, histograms, and Love plot (Ahmed et al. 2006; Cochran and Rubin 1973; Pan and Bai 2015a; Pattanayak 2015; Rosenbaum and Rubin 1985).

5 Issues in and debates about propensity score methods

As researchers embrace the advantages of using propensity score methods, some significant issues should be noted. It is also beneficial for the researchers to understand the concerns and debates about the use of these methods.

5.1 Issues in propensity score methods

The most conspicuous issue in propensity score methods is *hidden bias*. The fundamental theory of propensity score methods assumes that all the confounding variables are observable so that propensity scores calculated from the observed covariates can accurately represent the distributions of all confounding variables. Unfortunately, hidden bias due to unobservable confounding variables often exists because we cannot observe all confounding variables. To mitigate this issue, the selection of covariates should be first guided by theory, and then researchers should include all potential covariates that we could observe in propensity score estimation models (Pan and Bai 2016b). It is as important or more important to conduct sensitivity analysis for testing the model sensitivity to hidden bias from unobserved confounding variables.

Other important issues in propensity score methods are mainly related to propensity score matching such as matching with or without replacement or issues of sample reduction after matching. Matching with or without replacement should be considered each time when conducting propensity score matching, because the two different matching approaches are likely to produce significantly different matched data, especially when sample size is small (Pan and Bai 2015a). The selection of the two matching approaches also affects the treatment effect estimation after matching. For example, if matching with replacement is used, the analysis for estimating the treatment effect after matching should incorporate weighted scores so as to balance the subjects who appear multiple times in the matched data.

Propensity score matching usually removes unmatched subjects due to selection bias or covariate imbalance. Such sample reduction after matching may result in a matched sample unrepresentative of the target population. To combat this problem, large samples are preferable for implementing propensity score matching because large samples can produce more reliable results (Bai 2011b; Hirano et al. 2003; Månsson et al. 2007; Rubin 1997). Otherwise, different propensity score methods such as propensity score weighting should be considered.

5.2 Debates about propensity score methods

Just as many other statistical methods have been criticized, propensity score methods have also been questioned. There are two sides to the debates about propensity score methods (e.g., Pearl 2010; King and Nielsen 2016).

Pearl (2010) first argued that associational concepts can be defined in terms of joint distribution of observed variables such as the distribution of propensity scores, but causal concepts cannot be. Therefore, propensity scores would not be a causal concept, and only experimental control can verify causal assumptions. Pearl's argument seems correct, but it does not mean that propensity score methods ignore such an issue. In fact, propensity score methods have an assumption of the strongly ignorable treatment assignment that addresses the causal problem due to unobserved confounding variables. In practice, it is possible to estimate causality using observational data as long as all important confounding variables are well controlled. Furthermore, sensitivity analysis can test how sensitive of the treatment effect estimation with propensity scores to uncontrolled, but less influential confounding variables so as to safeguard the causal claims using propensity score methods. As Rubin (2009) pointed out, Pearl's argument might be irrelevant to propensity scores. Propensity score methods are intended to be outcome free, and the assumption of the strongly ignorable treatment is designed to be conditional on all observed values; therefore, it is assumed to control the influence of all possible confounding variables.

King and Nielsen (2016) hold a different opinion in regards to propensity score matching and argue that (a) propensity score matching cannot approximate a completely randomized experiment, (b) it is not comparable to a fully blocked randomized experiment, and (c) it is problematic due to some observations that increase imbalance and model dependency. These concerns seem reasonable when researchers overlook the assumptions of propensity score matching, which happens often. Thus, researchers are strongly urged to review the literature (e.g., King and Nielsen 2016; Pan and Bai 2016b; Rubin 2009) when implementing propensity score matching in their research.

6 Available software packages for propensity score methods

There is a variety of software packages available for implementing propensity score methods, including SAS, R, STATA, and SPSS. Some packages also have functions to combine propensity score procedures with treatment effect estimation. All the packages have advantages and disadvantages on specific propensity score methods.

For example, *MatchIt* in R (Ho et al. 2011) has most types of propensity score matching techniques including nearest neighbor matching, caliper matching, optimal matching, full matching, and genetic matching as well as subclassification. *MatchIt* also allows researchers to conduct matching with or without replacement and 1-to-1 or 1-to-many matching. The R package is easy to implement. Some STATA modules for propensity score methods (e.g., Leuven and Sianesi 2012) are also straightforward to use along with treatment effect estimation functions, which has an advantage over other packages. SAS (SAS Institute Inc. 2017a, b) recently developed two procedures for treatment effect estimation (PROC CAUSALTRT) and propensity score matching (PORC PSMATCH). The two procedures are available in SAS/STAT 14.2 or a newer version, or in the free SAS University Edition. SPSS modules for propensity score matching are available in its pull-down menu. In SPSS, we can also use Python-based extensions FUZZY to install add-on packages to implement *MatchIt* (Thoemmes 2012). Schuler (2015) provided a comprehensive survey on the use of all the software packages for propensity score methods along with useful code and examples.

7 Conclusion

Propensity score methods are popular and effective statistical techniques for reducing selection bias in observational data, and they increase the validity of causal inference based on observational studies. Some researchers have raised concerns about the rationale and applicability of propensity score methods. We addressed these concerns by reviewing the development history and the assumptions of propensity score methods, followed by the fundamental techniques of and software packages for propensity score methods. Our aim is to provide information about propensity score methods from a historical point of view, to emphasize the importance of checking assumptions, and to help researchers select the best methods for their observational studies.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ahmed A, Husain A, Love TE, Gambassi G, Dell'Italia LJ, Francis GS, Gheorghiade M, Allman RM, Meleth S, Bourge RC (2006) Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. Eur Heart J 27(12):1431–1439

Austin PC (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res 46(3):399–424

Austin PC, Stuart EA (2015) Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med 34(28):3661–3679

Bai H (2011a) A comparison of propensity score matching methods for reducing selection bias. Int J Res Method Educ 34(1):81–107

Bai H (2011b) Using propensity score analysis for making causal claims in research articles. Educ Psychol Rev 23:273–278

Bai H (2013) A bootstrap procedure of propensity score estimation. J Exp Educ 81(2):157–177

Bai H (2015) Methodological considerations in implementing propensity score matching. In: Pan W, Bai H (eds) Propensity score analysis: fundamentals, developments, and extensions. Guilford Press, New York, pp 74–88

Bang H, Robins JM (2005) Doubly robust estimation in missing data and causal inference models. Biometrics 61(4):962–973

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T (2006) Variable selection for propensity score models. Am J Epidemiol 163(12):1149–1156

Caliendo M, Kopeinig S (2008) Some practical guidance for the implementation of propensity score matching. J Econ Surveys 221:31–72

Cochran WG, Rubin DB (1973) Controlling bias in observational studies: a review. Sankhyā Indian J Stat Ser A 35(4):417–446

Dehejia RH, Wahba S (2002) Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat 84(1):151–161

Diamond A, Sekhon JS (2013) Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Rev Econ Stat 95:932–945

Fisher RA (1951) The design of experiments. Oliver & Boyd, Edinburgh

Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M (2001) Doubly robust estimation of causal effects. Am J Epidemiol 173(7):761–767

Greenland S (2005) Multiple-bias modelling for analysis of observational data. J R Stat Soc Ser A (Stat Soc) 168(2):267–306

Groenwold RHH, Nelson DB, Nichol KL, Hoes AW, Hak E (2010) Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. Int J Epidemiol 39(1):107–117

Guo S, Barth RP, Gibbons C (2006) Propensity score matching strategies for evaluating substance abuse services for child welfare clients. Child Youth Serv Rev 28(4):357–383

Hamilton MA (1979) Choosing the parameter for a 2×2 table or a 2×2×2 table analysis. Am J Epidemiol 109(3):362–375

Hansen BB (2004) Full matching in an observational study of coaching for the SAT. J Am Stat Assoc 99(467):609–618

Harder VS, Stuart EA, Anthony JC (2010) Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychol Methods 15(3):234–249

Heckman JJ, Ichimura H, Todd PE (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Rev Econ Stud 64(4):605–654

Hirano K, Imbens GW (2001) Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Serv Outcomes Res Method 2(3):259–278

Hirano K, Imbens GW, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71(4):1161–1189

Ho DE, Imai K, King G, Stuart EA (2011) MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw 42(8):1–28

Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81(396):945–960

Huesch MD (2013) External adjustment sensitivity analysis for unmeasured confounding: an application to coronary stent outcomes, Pennsylvania 2004–2008. Health Serv Res 48(3):1191–1214

Imai K, Ratkovic M (2014) Covariate balancing propensity score. J R Stat Soc Ser B (Stat Methodol) 76(1):243–263

Keele LJ (2015) Package ‘rbounds’, version 2.1. https://cran.r-project.org/web/packages/rbounds/rbounds.pdf. Accessed 20 Jan 2016

Kempthorne O (1952) The design and analysis of experiments. Wiley, Oxford

King G, Nielsen R (2016) Why propensity scores should not be used for matching. https://gking.harvard.edu/files/gking/files/psnot.pdf. Accessed 26 June 2017

Lemon SC, Roy JR, Clark MA, Friedmann PD, Rakowski WR (2003) Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Ann Behav Med 26:172–181

Leuven E, Sianesi B (2012) PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components S432001. Boston College Department of Economics. http://ideas.repec.org/c/boc/bocode/s432001.html. Accessed 6 May 2014

Li L, Shen C, Wu AC, Li X (2011) Propensity score-based sensitivity analysis method for uncontrolled confounding. Am J Epidemiol 174(3):345–353

Li J, Handorf E, Bekelman J, Mitra N (2017) Propensity score and doubly robust methods for estimating the effect of treatment on censored cost. Stat Med 35(12):1985–1999

Lin DY, Psaty BM, Kronmal RA (1998) Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 54(3):948–963

MacLehose RF, Kaufman S, Kaufman JS, Poole C (2005) Bounding causal effects under uncontrolled confounding using counterfactuals. Epidemiology 16(4):548–555

Månsson R, Joffe MM, Sun W, Hennessy S (2007) On the estimation and use of propensity scores in case-control and case-cohort studies. Am J Epidemiol 166(3):332–339

McCaffrey DF, Ridgeway G, Morral AR (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 9(4):403–425

Pan W, Bai H (eds) (2015a) Propensity score analysis: fundamentals and developments. Guilford Press, New York

Pan W, Bai H (2015b) Propensity score interval matching: using bootstrap confidence intervals for accommodating estimation errors of propensity scores. BMC Med Res Methodol 15(53):1–9

Pan W, Bai H (2016a) A robustness index of propensity score estimation to uncontrolled confounders. In: He H, Wu P, Chen D (eds) Statistical causal inferences and their applications in public health research. Springer, New York, pp 91–100

Pan W, Bai H (2016b) Propensity score methods in nursing research: take advantage of them but proceed with caution. Nurs Res 65(6):421–424

Pattanayak CW (2015) Evaluating covariate balance. In: Pan W, Bai H (eds) Propensity score analysis: fundamentals and developments. Guilford Press, New York, pp 89–112

Pearl J (2010) The foundations of causal inference. Sociol Methodol 40(1):75–149

Robins JM, Hernan MA, Brumback B (2000a) Marginal structural models and causal inference in epidemiology. Epidemiology 11:550–560

Robins JM, Rotnitzky A, Scharfstein DO (2000b) Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME, Berry D (eds) Statistical models in epidemiology, the environment, and clinical trials. Springer, New York, pp 1–94

Rosenbaum PR (1989) Optimal matching for observational studies. J Am Stat Assoc 84(408):1024–1032

Rosenbaum PR (2010) Observational studies, 2nd edn. Springer, New York

Rosenbaum PR, Rubin DB (1983a) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. J R Stat Soc Ser B (Methodol) 45(2):212–218

Rosenbaum PR, Rubin DB (1983b) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55

Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 79(387):516–524

Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 39(1):33–38

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66(5):688

Rubin DB (1977) Assignment to treatment group on the basis of a covariate. J Educ Behav Stat 2(1):1–26

Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann Stat 6(1):34–58

Rubin DB (1980) Bias reduction using Mahalanobis metric matching. Biometrics 36:293–298

Rubin DB (1997) Estimating causal effects from large data sets using propensity scores. Ann Intern Med 127(8_Part_2):757–763

Rubin DB (2001) Using propensity scores to help design observational studies: application to the tobacco litigation. Health Serv Outcomes Res Method 2(3–4):169–188

Rubin DB (2008) For objective causal inference, design trumps analysis. Ann Appl Stat 2(3):808–840

Rubin DB (2009) Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? Stat Med 28(9):1420–1423

Rubin DB, Thomas N (1996) Matching using estimated propensity scores: relating theory to practice. Biometrics 52:249–264. https://doi.org/10.2307/2533160

SAS Institute Inc. (2017a) SAS/STAT® 14.3 user's guide: the CAUSALTRT procedure. SAS Institute Inc., Cary, NC

SAS Institute Inc. (2017b) SAS/STAT® 14.3 user's guide: the PSMATCH procedure. SAS Institute Inc., Cary, NC

Schafer JL, Kang J (2008) Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol Methods 13:279–313

Schneeweiss S (2006) Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf 15(5):291–303

Schuler M (2015) Overview of implementing propensity score analyses in statistical software. In: Pan W, Bai H (eds) Propensity score analysis: fundamentals and developments. Guilford Press, New York, pp 20–48

Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin, Boston

Shadish WR, Clark MH, Steiner PM (2008) Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. J Am Stat Assoc 3(484):1334–1344

Smith JA, Todd PE (2005) Does matching overcome LaLonde's critique of nonexperimental estimators? J Econ 125:305–353

Stone CA, Tang Y (2013) Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. Pract Assess Res Eval 18(13):1–12

Thoemmes F (2012) Propensity score matching in SPSS. https://arxiv.org/abs/1201.6385. Accessed 26 May 2014

Westreich D, Lessler J, Funk MJ (2010) Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J Clin Epidemiol 36(8):826–833

Winship C, Morgan SL (1999) The estimation of causal effects from observational data. Ann Rev Sociol 25:659–706